



ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under Resourced Areas of Machine Translation

www accurat-project.eu

Project no. 248347

Deliverable D5.4

**Report on requirements, implementation and
evaluation of usability in application for software
localization**

Version No. 1.0

29/06/2012

Document Information

Deliverable number:	D5.4
Deliverable title:	Report on requirements, implementation and evaluation of usability in application for software localization
Due date of deliverable:	30/06/2012
Actual submission date of deliverable:	29/06/2012
Main Author(s):	Mārcis Pinnis, Inguna Skadiņa, Andrejs Vasiļjevs
Participants:	Tilde
Internal reviewer:	Tilde
Workpackage:	WP5
Workpackage title:	Evaluation of usability in applications
Workpackage leader:	Tilde
Dissemination Level:	PU : public
Version:	V1.0
Keywords:	Software localization, translation memories, comparable corpora

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
V0.1	10/05/2012	Draft	Tilde	Fishbone	Fishbone approved
V0.2		First draft	Tilde	First draft	First draft approved
V1.0	30/06/2012	Final version	Tilde	Major improvements	Submitted to PO

EXECUTIVE SUMMARY

This deliverable describes adaptation and evaluation of ACCURAT statistical machine translation (SMT) system for software localization task. The ACCURAT baseline SMT system developed in WP4 was enriched with data extracted from IT domain comparable corpus using ACCURAT tools. The developed SMT system was used to assess translation quality and usability of ACCURAT methodology for this particular task.

Machine translation output quality was evaluated using automated metrics (BLEU) and human evaluation (system comparison and software localization workflow). For software localization task ACCURAT IT domain SMT system was integrated with Trados CAT tool and applied in real life software localization.

Table of Contents

Abbreviations.....	4
Introduction.....	5
1. Localization Requirements for MT.....	6
2. Related work.....	8
3. Application of ACCURAT tools in the localisation process.....	9
4. Evaluation Objects: software localisation domain tuned SMT system.....	10
4.1. Preparing comparable corpora.....	10
4.2. Extracting data from comparable corpora.....	10
4.2.1. Parallel sentence pair extraction.....	10
4.2.2. Extraction of translated term pairs.....	11
4.3. Training SMT systems.....	13
4.3.1. Baseline SMT System.....	13
4.3.2. Intermediate ACCURAT improved SMT system.....	14
4.3.3. ACCURAT improved SMT system.....	17
5. Evaluation Methodology.....	17
6. Automatic evaluation.....	18
7. Comparative evaluation.....	18
7.1. Evaluation environment.....	18
7.2. Evaluation methodology.....	19
7.3. Evaluation results.....	20
8. Evaluation in Localization Scenario.....	21
8.1. Evaluation methodology.....	21
8.2. Test data.....	21
8.3. Evaluation scenarios.....	21
8.4. Tools: SDL Trados + SMT.....	22
8.5. Evaluation and results.....	23
8.5.1. Test execution.....	23
8.5.2. Evaluation procedure.....	23
8.5.3. Evaluation results.....	24
Conclusion.....	26
References.....	27
Appendix 1. Creation of English-Latvian comparable corpora.....	28
Appendix 2. Tilde Translation Quality Assessment Form.....	31
Appendix 3. Detailed evaluation results in localisation scenario.....	35

Abbreviations

Abbreviation	Term/definition
BLEU	Bilingual Evaluation Understudy
CAT	Computer assisted translation
FMC	Focused monolingual crawler
IT	Information technology
MERT	Minimum error rate training
METEOR	Metric for evaluation of translation with explicit ordering
MT	Machine translation
NIST Metric	Metric of the National Institute of Standards and Technology
QA	Quality assurance
SMT	Statistical machine translation
SOAP	Simple object access protocol
SW	Software
TEA	Terminology Aligner
TER	Translation error rate
TM	Translation memory
TWSC	Tilde's Wrapper System for CollTerm

Introduction

The goal of WP5 is to adapt and validate the ACCURAT tools and methodology in practical applications and assess usability and translation quality of MT elaborated with data from comparable corpora in these applications. This deliverable focuses on application of ACCURAT results in the software localization process.

The first section discusses key expectations and requirements of localization industry regarding machine translation systems that are relevant to the ACCURAT project. In recent years, the localization industry has started using MT to reduce human workload in the translation business. However, until now due to the constraints of under-resourced areas MT application has been limited to larger languages and major domains only. This deliverable describes adaptation and evaluation of ACCURAT methods in software localization from English into under-resourced languages.

The second section provides a brief overview of related work in MT applicability for real world localization tasks.

In our work that is described in the following sections we tested feasibility of the ACCURAT MT solution in software localization process from English into Latvian. For this task SMT system trained on publicly available parallel data was used as a baseline. Development of the baseline system is described in the Section 3.1. Section 3.2 describes adaptation of system to the IT software domain using in-domain comparable corpus and software domain data collected from the Web with ACCURAT FMC Tool¹.

In general we followed the evaluation scenario described in Deliverable 5.1 *Evaluation plan*. At first baseline and ACCURAT improved SMT systems were evaluated and compared using automatic evaluation metrics BLEU (Papineni et al., 2002). It clearly shows that in the software domain the SMT system enriched with in-domain comparable data performs much better than the baseline system. This evaluation is described in the Section 5.

Purpose of the next evaluation described in the Section 7 is to assess whether ACCURAT improved SMT system can help to increase translators' productivity in real-world software localization work. Taking into account low quality of the baseline system in software domain, we decided not to use it in this evaluation as it would degrade translators' performance. Instead translators' productivity was measured in two scenarios: (i) in a usual setup using only suggestions from translation memories and (ii) providing MT suggestions in addition to the suggestions from translation memory. In the second scenario translation suggestions from MT were provided for those translation segments that did not have an exact match or a close match in the translation memory. Translators' productivity was calculated as a number of words translated per hour.

We also evaluated impact that the proposed solution has on the translation quality. Quality of the final translation output was evaluated using the standard internal quality assessment procedure.

In this work we cooperated with the ICT-PSP project LetsMT!. SMT systems described in this deliverable were trained and run on the LetsMT! platform². It is an online platform for

¹ Web data was collected by the project partner ILSP. For FMC applied methods refer to the Deliverable D3.4 "Report on methods for collection of comparable corpora". For the software documentation of FMC refer to the Deliverable D3. "Tools for building comparable corpus from the Web".

² <http://www.letsmt.eu>

sharing of training data and building user tailored machine translation systems (Vasiljevs et al., 2010). The LetsMT! plug-in was used for integrating MT into translators' workbench SDL Trados 2009 that was used in this evaluation.

1. Localization Requirements for MT

Localization industry is dealing with translation and cultural adaptation of software, websites and IT products for other markets and language communities. It is a major part of a more general language service sector. According to the data from the Common Sense Advisory (Kelly et al., 2012) there are more than 26 000 companies with two or more employees operating in the language service industry. Common Sense Advisory calculates that the market for outsourced language services is worth US\$33.523 billion in 2012 and it is growing at an annual rate of 12.17%. Europe (49.38%) makes up the largest region for language services, followed by North America (34.85%) and Asia (12.88%).

This industry experiences a growing pressure on efficiency and performance, especially due to the fact that volumes of texts that need to be translated are growing at a greater rate than the availability of human translation, and translation results are expected in real-time. Industry experiences a strong shift from larger translation projects with delivery time in several months to a huge stream of small translation requests to be fulfilled in a day or few.

The key forces driving language service market are (WinterGreen Research, 2011):

- Increase efficiency;
- Accelerate translations;
- Allow more projects to be accepted;
- Grow revenues;
- Reduce translation costs;
- Leverage company's multilingual resources.

Translation memories (TM) are widely used in localization industry to increase translators' productivity and consistency of translated material. Translation memories can significantly improve the efficiency of localization if the new text is similar to the previously translated material. However, if the text is in a different domain than the TM or in the same domain from a different customer using different terminology, support from the TM is minimal.

These factors drive a growing awareness and interest of localization industry in application of machine translation to increase volumes of translation and decrease costs of the service.

For the development of MT in the localization and translation industry, huge pools of parallel texts in a variety of industry formats have been accumulated, but the use of this data alone does not fully utilize the benefits of modern MT technology.

Machine translation can be used in several scenarios relevant for the localization services:

1) To serve as an additional source of reference for human translators.

In this scenario human translator works with traditional computer aided translation tools (CAT) and occasionally uses MT systems like Google Translator to translate a phrase or sentence. This is a quite typical usage of MT in companies that have not yet started organized MT implementation.

2) To provide MT translation suggestions for source strings without a full-match or close match in the translation memory.

This scenario requires integration of an MT system within a CAT tool to enable a live translation of respective strings (or integration of MT into the source text preprocessing workflow to provide pre-translated MT suggestions). In this scenario a translator receives MT suggestions as a supplement to the suggestions provided

from translation memory. The translator is in a full control to choose whether to use the MT suggestion as it is, to take it and make necessary post-editing, or to ignore a bad suggestion and do translation from scratch.

3) To provide a full translation of source text for post editing by a human translator.

In this scenario machine translation is provided for a full text or document. Instead of sentence by sentence translation of the source text, the translator receives a machine translated target document and makes necessary post edits in it.

4) To provide full translation of the source text to be used “as is”.

In this scenario machine translation output is used without post-editing by human translator/reviewer. It is applicable for translations used only for gisting purposes or for highly standardized content where machine translation can provide quality that is on par with a human translation quality.

5) To provide an instant on-demand machine translation for websites or client’s solutions.

It is applicable for rapidly changing content where translation quality is less critical than enabling access to the general content of material in another language.

6) To provide an instant on-demand machine translation with an option to edit a translation and make dynamic adaptations.

This scenario is in-between the instant full MT and the post-editing scenario. Here the user has the possibility to post-edit an on-demand translation provided by MT system. The system “learns” and improves from these post edits.

The following are some of the key requirements for application of machine translation in the localization industry that are relevant to the ACCURAT project:

- *Quality of translation*
Quality continues to be the major concern in application of MT for professional services. To apply MT in the instant translation scenarios its comprehension, accuracy and fluency should be sufficient to meet the basic user needs that depend on the type of content or service. For the post-editing scenarios MT output quality should be sufficient to motivate post-editing efforts versus human translation from scratch.
- *Language coverage*
Availability and quality of MT systems vary dramatically between languages. Smaller languages are often not supported by commercial MT providers or translation quality is match worse comparing to the translation between larger languages.
- *Domain coverage*
Domain specific MT systems usually perform much better for in-domain translation comparing to general MT systems. Still only few domains in few translation directions are served by commercial MT offerings.
- *Terminology usage*
Usage of domain and project specific terminology is one of the key requirements for translation quality. Rule-based MT systems can be adapted by providing terminology dictionaries. For SMT systems currently there is no commercial offering available for terminology adaptation.
- *Cost of adaptation*
Adaptation of MT to the domain, terminology requirements or other needs of a particular client is offered as an individual service that is too expensive for the majority of language service providers.

In the ACCURAT project we focus on the application of ACCURAT tools for the MT application scenario where MT is integrated into a CAT tool to provide translation suggestions.

Increasing the efficiency of translation process without degradation of quality is the most important goal for a localization service provider. Efficiency of translation process directly depends on the performance of translators. Performance is usually measured in the number of words translated per hour. In this deliverable we describe application of ACCURAT methods to create improved MT systems that can boost translators' productivity enabling them to increase the speed of translation.

2. Related work

Although the idea to use MT in the localization process is not new, it has got more attention from researchers and localization industry only recently.

Different aspects of post-editing and machine translatability have been researched since the 90-ies. A comprehensive overview of research on machine translatability and post-editing has been provided by O'Brien (2005). However this work mainly concentrates on the cognitive aspects, not so much on productivity in the localization industry.

Recently several productivity tests have been performed in translation and localization industry settings at Microsoft (Schmidtke, 2008), Adobe (Flournoy and Duran, 2009), Autodesk (Plitt and Masselot, 2010) and Tilde (Skadiņš et al., 2011).

The Microsoft Research trained SMT on MS tech domain was used for 3 languages for Office Online 2007 localization: Spanish, French and German. By applying MT to all new words on average a 5-10% productivity improvement was gained.

Adobe performed two experiments. At first a small test set of 800-2000 words was machine translated and post-edited. Then, based on the positive results, about 200,000 words of new text were localized. The rule-based MT was used for translation into Russian (PROMT) and SMT for Spanish and French (Language Weaver). For the first experiment authors reported the speed-up between 22% and 51%, while for the second experiment authors reported preliminary results: "the MT post-editing was performed 40% to 45% faster than human translation for comparable text".

At Autodesk, a Moses SMT system was evaluated for translation from English to French, Italian, German and Spanish by three translators for each language pair. To measure translation time a special workbench was designed to capture keyboard and pause times for each sentence. Authors reported that although by using MT all translators worked faster, it was in varying proportions: from 20% to 131%. They concluded that MT allowed translators to improve their throughput on average by 74%. They also reported that optimum throughput has been reached for sentences of around 25 words in length.

Tilde performed an experiment on the application of English-Latvian SMT in localization through the integration of MT into the SDL Trados 2009 translation environment. The SMT system was trained with Moses SMT toolkit (Koehn et al., 2007). In this experiment performance of a translator translating with translation memory (TM) only and with combination of TM and MT was measured as well as a quality assessment for texts was performed according to the Tilde's standard internal quality assessment procedure. Five translators with different levels of experience and average performance were involved in the evaluation. Documents (950-1050 adjusted words each) for translation were selected from the incoming work pipeline and split in half. The first part of the document was translated with TM and the second half of the document – using the SMT and TM. Altogether 54 documents were translated. In this experiment usage of MT suggestions in addition to the use of the

translation memories increased productivity of the translators in average from 550 to 731 words per hour (32.9% improvement). However, there were significant performance differences in the various translation tasks; the standard deviation of productivity in the baseline and MT scenarios were 213.8 and 315.5 respectively. At the same time the error score increased for all translators. The total increase in the error score was from 20.2 to 28.6 points, which according to the internal quality assessment scale still remained at the quality evaluation grade “Good”.

As industry experiences a growing pressure on efficiency and performance, some developers have already integrated MT in their products or provide such solutions for MT developers. For instance, SDL Trados Studio 2009 supports 3 machine translation engines: SDL Enterprise Translation Server, Language Weaver, and Google Translate. ESteam TRANSLATOR and Kilgrey’s memoQ are other systems providing integration of MT.

3. Application of ACCURAT tools in the localisation process

Methods and tools developed within the ACCURAT project can be used in the localisation process for SMT system adaptation purposes. In situations where parallel data for good quality SMT system training is not available ACCURAT methods allow finding in-domain comparable corpora and extracting parallel sentences and in-domain terminology that can be used to adapt broad domain SMT systems to specific domains. This deliverable shows one such use case of ACCURAT tools and methods for software localisation purposes.

The usual process chain of system adaptation with ACCURAT results involves:

- Collection of in-domain comparable corpora. The ACCURAT *Focussed Monolingual Crawler (FMC)*; developed by ILSP) allows crawling the Web using in-domain translated terms and seed URL lists and can be used for this task. After this step, we have additional training data for in-domain SMT language models.
- Alignment of in-domain comparable corpora. The ACCURAT comparability metrics tools *DictMetric* and *ComMetric* (developed by CTS) can be used to align the collected comparable corpora in the document level.
- Extraction of in-domain translated sentence pairs. The ACCURAT tool *LEXACC* (developed by RACAI) allows finding pseudo-parallel sentence pairs within in-domain comparable corpora. After this step (see section), we have additional training data (sentence pairs) for in-domain SMT translation models.
- Extraction of in-domain translated term pairs. The ACCURAT tools *Tilde’s Wrapper System for CollTerm (TWSC)*, developed by TILDE and FFZG) and *TerminologyExtraction* (developed by RACAI) can be used for tagging of terms in each monolingual corpora and the ACCURAT tools *TerminologyAligner (TEA)*; developed by RACAI) and *MapperUSFD* (developed by USFD) can then be used for cross-lingual mapping of terms using the document alignment information acquired from one of the comparability metrics tools. After this step, we also have additional training data for in-domain training data (term pairs) for in-domain SMT translation models.
- Training the intermediate ACCURAT improved SMT system (see Section 4.3.2). We use all available parallel data (including the extracted data from the comparable corpora) and all available monolingual data to train the SMT system. We build one translation model and two language models - a general domain

language model and an in-domain language model using the acquired in-domain corpus from the first step.

- Using extracted bilingual terminology to adapt the translation model of the SMT system. We use the extracted in-domain term pairs in order to transform the phrase table of the SMT system's translation model to a term-aware phrase table (see section 4.3.3).
- Integrating the SMT system within a CAT system, for instance, the SDL Trados 2009 translation environment.
- Using the SMT system in everyday localisation tasks.

4. Evaluation Objects: software localisation domain tuned SMT system

In this experiment we compare two SMT systems: a baseline system, trained on available parallel data, and a domain-adapted SMT system, trained on strongly comparable IT domain data. Both systems have been trained using the *LetsMT!* platform which provides integration with the SDL Trados 2009 translation environment that is actively used by translators involved in localisation. In the adaptation process we present also an intermediate system to show differences between two distinct SMT system adaptation steps.

The adaptation process follows the process chain described in the previous section. The further sections describe each of the steps in more detail.

4.1. Preparing comparable corpora

In this experiment we assume that the user already has access to strongly comparable in-domain corpora, thus we do not describe how to collect comparable corpora from various sources.

For our task we acquired four different software localisation domain English-Latvian comparable corpora, three of which have been created from software manuals ("*SW Manual {1/2/3}*") and one is a combination of web crawled corpora and software manuals ("*SW Mixed*"). For a detailed description of how these corpora were created refer to Appendix 1. Although, these comparable corpora have been artificially created, the whole process chain of system adaptation is the same for any comparable corpus such as corpora automatically acquired from the Web.

4.2. Extracting data from comparable corpora

After acquiring the comparable corpora, possibly parallel (translated) data for SMT was extracted. Two types of possibly parallel data were extracted from the comparable corpora – parallel sentence pairs and translated term pairs.

4.2.1. Parallel sentence pair extraction

The parallel sentence extractor *LEXACC* (developed by RACAI) described in Deliverable 2.6. *Toolkit for multi-level alignment and information extraction from comparable corpora* was used to extract parallel sentences. All corpora were pre-processed before parallel data extraction – texts in both languages were broken into sentences (one sentence per line) and tokenized (tokens were separated by a space). The statistics of the extracted pseudo-parallel (parallel, strongly comparable and weakly comparable) data with *LEXACC* are summarised in Table 1.

Table 1. Sentence pairs extracted with LEXACC

Corpus	Threshold of LEXACC confidence score	Unique sentence pairs	% of all unique sentences		Tokens in unique sentence pairs	
			English	Latvian	English	Latvian
SW Manual 1	0.1	28,778	14.57 %	16.03 %	439,328	419,797
	0.6	1,308	0.66 %	0.73 %	14,478	13,859
SW Manual 2	0.1	37,274	20.80 %	18.54 %	555,203	617,179
	0.6	1,517	0.85 %	0.75 %	16,886	17,582
SW Manual 3	0.1	44,294	39.55 %	39.08 %	877,619	875,497
	0.6	6,895	6.16 %	6.08 %	148,067	136,944
SW Mixed	0.1	615,489	42.93 %	48.32 %	10,771,631	9,505,609
	0.35	561,994	39.56 %	45.12 %	9,608,942	8,387,031

Because every corpus is different (in terms of comparable data distribution and the comparability levels) different *LEXACC* confidence score thresholds were applied. Table 1 shows information about extracted data using two different thresholds. The results with the threshold 0.1 are given as reference for the total number of sentence pairs that can be extracted from each corpus with *LEXACC*. The second threshold was selected by manual result inspection so that most (more than 90%) of the extracted sentence pairs would be strongly comparable and parallel. Only the sentence pairs extracted with the manually set threshold were used for the SMT system adaptation.

From the table we can see that the first two corpora created from different versions of software manuals did not have many overlapping content. Results also show that *LEXACC* is able to find from approximately 40% to 50% of parallel sentences in the *SW Mixed* corpus for the English-Latvian language pair.

4.2.2. Extraction of translated term pairs

Terms are important domain specific resource and integration of terms within SMT systems has an important role in the SMT adaptation process to software localisation domains. ACCURAT methods allow acquiring in-domain term pairs (Pinnis et al., 2012b) from comparable corpora, which can then be used in the adaptation process. The overview of term extraction process is presented in Figure 1.

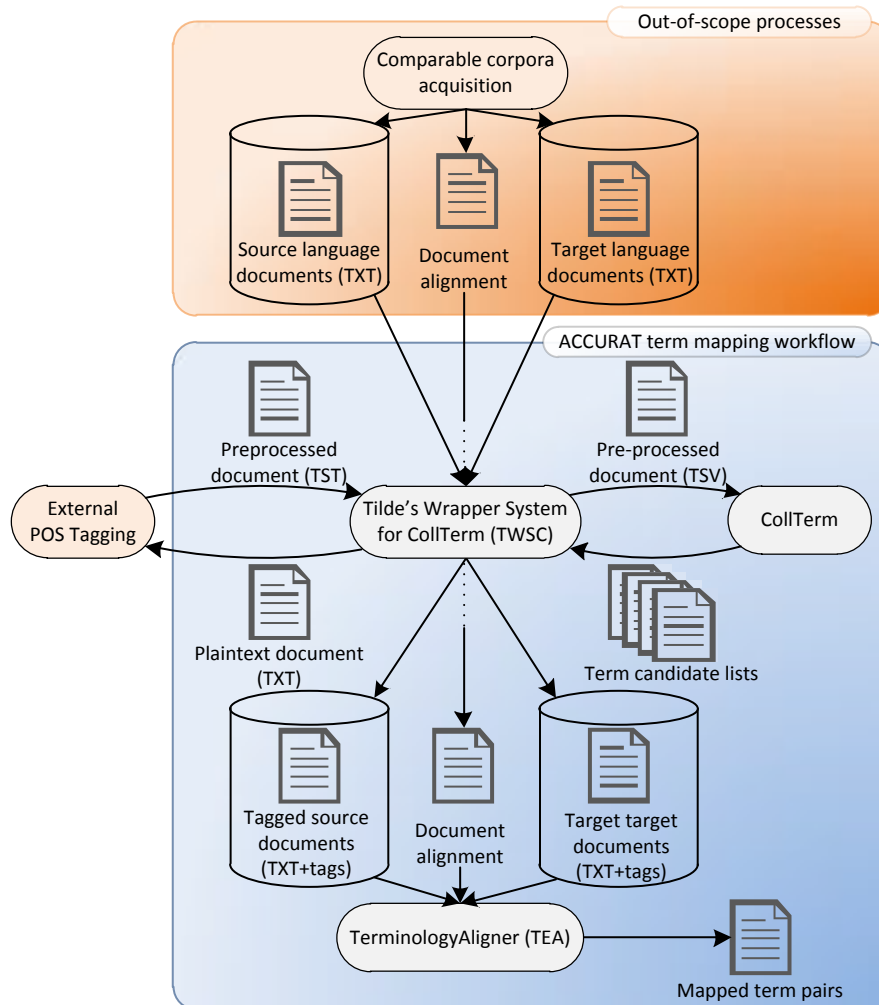


Figure 1. Extraction of translated term pairs

All comparable corpora were monolingually tagged with *TWSC* (Pinnis et al. 2012a). For Latvian a configuration that maximises the tagging F-measure was used. For English a default configuration was used as *TWSC* has not been fine-tuned for English term tagging.

After monolingual term tagging, *TEA* was used for term mapping on the document pairs acquired in the corpora acquisition phase. *TEA* creates translated term pairs with a translation confidence score. For mapping, a confidence score threshold of 0.7 has been used (to achieve precision of about 90%).

In the term tagging step we try to achieve a relatively high recall, because the cross-language mapping with *TEA* allows filtering out most of the wrong terms. It has been shown by (Pinnis et al. 2012b) that *TWSC* achieves a precision of 50% to 53%. This means that many of the tagged phrases are not actual terms. However, even with a relatively low tagging precision it is possible to achieve a high cross-lingual mapping precision. The statistics of both monolingual corpora and the mapped terms are given in Table 2.

Table 2. Term tagging and mapping statistics

Corpus	Unique monolingual terms		Mapped term pairs	
	English	Latvian	Before filtering	After filtering
SW Manual 1	56,265	75,710	235	180
SW Manual 2	34,313	120,746	250	218
SW Manual 3	36,838	74,971	362	291
SW Mixed	415,401	2,566,891	3,501	3,393

The translated term pairs were further filtered so that for each Latvian term only the term pairs with high *TEA* translation confidence score are preserved. We used Latvian term translations to filter term pairs, because Latvian is a morphologically richer language and multiple inflective forms of a word in most cases correspond to single English word form (although this is a “*rude*” filter, it increases the precision of term mapping to well over 90%).

Table 2 shows a big difference between monolingual and mapped terms, which is caused by restrictions of the mapper. However, the amount of mapped terms is enough to allow SMT system adaptation. It should also be noted that in our adaptation scenario (described further) translated single-word terms are more important than multi-word terms as the adaptation process of single-word terms partially covers also the multi-word pairs that have been missed by *TEA*.

4.3. Training SMT systems

Once all required data (large general domain parallel corpora, in-domain comparable corpora, in-domain pseudo-parallel sentence pairs and in-domain translated term pairs) is acquired, the SMT system training can be started.

4.3.1. Baseline SMT System

For the English-Latvian baseline system the DGT-TM³ parallel corpora of both releases (2007 and 2011) were used. The statistics of DGT-TM is shown in Table 3 below.

Table 3. Statistics of DGT-TM parallel corpora for English-Latvian

Corpus (version)	Sentence pairs (total)
DGT-TM (2007)	1,120,835
DGT-TM (2011)	1,753,983

Because of format limitations 94.31% of the DGT-TM (2011) was used for training (the original corpus contains 1,859,781 sentence pairs).

The parallel corpus was cleaned (filtered) in the *LetsMT!* platform in order to remove corrupt sentence pairs. Sentence pairs were considered corrupt if:

- The sentence pair contained words longer than 50 characters (captures mistakes where words have been written together).
- Either of sentences in a sentence pair was longer than 1000 characters (captures sentence breaking errors).
- Three or more capitalized words were joined together (for instance, “*TheEuropeanUnion*”; captures mistakes where words have been written together).

³ The DGT Multilingual Translation Memory of the Acquis Communautaire: DGT-TM; available online at <http://langtech.jrc.it/DGT-TM.html>.

- Either of sentences in a sentence pair contained special characters from a different language (for instance, if an English sentence would contain the Latvian special character “ā”, the corresponding sentence pair would be filtered out).
- Either of sentences contained a sequence of 5 or more space-separated characters (for instance, instead of the word “translate”, the English sentence would contain “t r a n s l a t e”).
- In either of sentences more than 36% of non-whitespace characters were digits (indicates of a code sequence rather than a sentence).
- In either of sentences less than 65% of non-whitespace characters were alphanumeric characters (digits and letters; this indicates of a code sequence rather than a sentence).
- A sentence in one language contained three times or more tokens than in the other language (indicates wrong sentence boundaries).
- Both sentences in a sentence pair were identical (indicates of a code segment or a non-translated text segment).

The DGT-TM (2007 and 2011 combined) English-Latvian parallel corpus contained in total 167,435 corrupt sentence pairs. After cleaning also 879,066 duplicate sentence pairs were removed. As a result, for training of the baseline system in the LetsMT! platform a total of **1,828,317 unique parallel sentence pairs** were used for translation model training.

The Latvian part of DGT-TM (2007 and 2011 combined) was used for language model training. After cleaning and duplicate removal the corpus contained a total of **1,736,384 unique Latvian sentences**.

The baseline system has been tuned with MERT using in-domain (software localisation domain) tuning data. The tuning data has been randomly extracted from proprietary software localisation translation memories (containing more than a million sentence pairs from software manuals and user interface texts) owned by Tilde. **The tuning set contains 1,837 unique sentence pairs** that have been manually edited by human translators. Tuning of the baseline system was done in 11 MERT iterations.

4.3.2. Intermediate ACCURAT improved SMT system

The adaptation to the software localisation domain was performed on the LetsMT! platform. The monolingual corpora and the parallel data (sentence pairs and term pairs) were uploaded to the LetsMT! platform and new system training was initiated. The overview of the training process is presented in Figure 3.

In order to create the translation model of the SMT system, the extracted in-domain parallel data (including sentence pairs and term pairs) was added to the existing parallel corpora (DGT-TM 2007 and 2011). The whole parallel corpus was then cleaned and filtered with the same techniques as for the baseline system. Filtering also involves removal of all parallel sentence pairs from the training data that occur in the tuning or evaluation sets.

The final statistics of the filtered corpora used in SMT training of the improved systems (intermediate and final) is as follows:

- **Parallel corpora** consists of:
 - 1,828,317 unique sentence pairs from the DGT-TM (2007 and 2011) corpora;
 - 558,168 unique sentence pairs from the in-domain LEXACC extracted corpora;

- 3,594 unique term pairs from the in-domain TEA extracted (and post-filtered) term pairs;
- **Monolingual corpora** consists of:
 - 1,576,623 unique general domain sentences from the DGT-TM (2007 and 2011) corpora;
 - 1,317,298 unique in-domain sentences from the acquired in-domain comparable corpora and the list of Latvian terms from the TEA extracted term pairs.

From this we can conclude that there has been some sentence pair overlap between the DGT-TM corpora (more precisely – 3,826 unique sentence pairs were found in both corpora) and the LEXACC extracted sentence pairs. This was expected as DGT-TM covers a broad domain and may also contain documents related to the IT domain. For language modelling, however, the sentences that overlap in in-domain and general domain monolingual corpora have been filtered out from the general domain monolingual corpus. Therefore, the DGT-TM monolingual corpus statistics between the baseline system and the adapted system do not match.

After filtering, for the intermediate (and also final) ACCURAT improved SMT system a translation model was trained from all available parallel data and two separate language models were trained from the monolingual corpora:

- The DGT-TM corpora corresponding Latvian sentences were used to build the general domain language model;
- The in-domain corpora acquired in the previous steps was used to build the in-domain language model.

At this step it is not yet possible to distinguish in-domain translation candidates from general domain translation candidates in the translation model. Therefore, we refer to this system as the Intermediate ACCURAT improved SMT system. To have comparative results with the fully adapted SMT system, we tuned the intermediate system with MERT using the in-domain tuning set of 1,837 unique in-domain sentence pairs. With this step we also wanted to see how big impact over the baseline system can be achieved by only adding an in-domain language model and additional in-domain parallel data for the translation model. The results of the intermediate adapted system are shown in Figure 3 and Table 4.

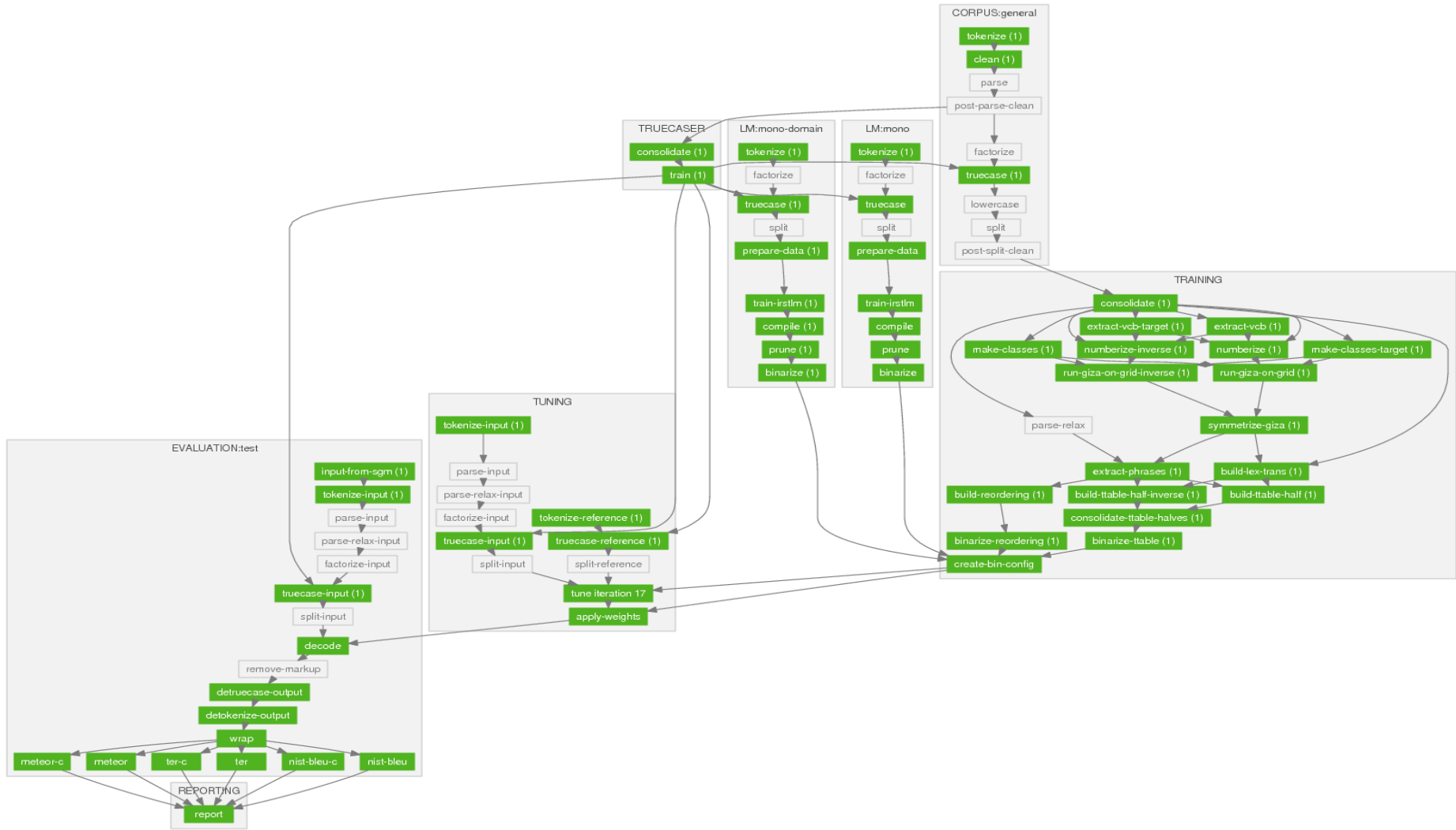


Figure 2. Training chart of the intermediate ACCURAT improved SMT system

Let's MT! beta Systems Corpora Translate Tools About Administration Welcome, Tildes lietotājs! Sign Out
Work in Tilde

Systems Create system

Source language: English Target language: Latvian Name / Title: Accurat.IT.Sys3 - add filter -

Status	Name / Title	Source language	Target language	Subject Domain	Permissions
● Not Started	Accurat.IT.Sys3	English	Latvian	Other	Private

Description: Accurat.IT.Sys3 - P: DGT (2007+2011) + LEXACC extracted data from artificial comparable corpora V3 + TEA3 terms V3
BLEU Score: 56.28
NIST Score: 9.1805
BLEU Score (CS): 54.81
NIST Score (CS): 8.9349
Monolingual corpus: 1 576 623 sentences
Parallel corpus: 2 390 079 sentences
Evaluation set: 926 sentences
Tuning set: 1 837 sentences
In-domain monolingual corpus: 1 317 298 sentences
Date Created: 2012.06.03 05:58:47 (UTC)

Date Modified: 2012.06.03 07:02:13 (UTC)
Date Accessed: 2012.06.03 07:02:13 (UTC)
Training started: 2012.06.05 08:26:16 (UTC)
Training finished: 2012.06.06 15:59:33 (UTC)
Creator (group): tilde
Owner (user): tilde
Parallel corpora: Accurat.IT.LEXACC_PART_1_(0.6),DGT-TM-2007,Accurat.IT.LEXACC_PART_2_(0.6),Accurat.IT.TEA_TERMS_PART_1_(0.7),Accurat.IT.LEXACC_PART_3_(0.35),DGT-TM-2011,Accurat.IT.TEA_TERMS_PART_3,Accurat.IT.TEA_TERMS_PART_2_(0.7)
Monolingual corpora: DGT-TM-2007,DGT-TM-2011
Corpora tuning: Accurat.IT.DEVELOPMENT_(TUNING)
Corpora evaluation: Accurat.IT.EVAL
In-domain monolingual corpora: Accurat.IT.LV_mono_corpus_part_2,Accurat.IT.LEXACC_PART_1_(0.6),Accurat.IT.TEA_TERMS_PART_1_(0.7),Accurat.IT.TEA_TERMS_PART_2_(0.7),Accurat.IT.LV_mono_corpus_part_3,Accurat.IT.LEXACC_PART_2_(0.6),Accurat.IT.LEXACC_PART_3_(0.35),Accurat.IT.TEA_TERMS_PART_3

Details Start instance View training chart

This is a list of public and private statistical machine translation (SMT) systems.
An SMT system is an automatic text translator you can build and train to translate texts in general or specific subjects (domains).

Figure 3. General characteristics of the intermediate ACCURAT improved SMT system

4.3.3. ACCURAT improved SMT system

To make in-domain translation candidates distinguishable from general domain translation candidates, the phrase table of the intermediate ACCURAT improved SMT system was further transformed to a term-aware phrase table. This means that a sixth feature was added to the default 5 features that are used in Moses phrase tables. This sixth feature received the following values:

- “1” if a phrase in both languages did not contain a term pair from the filtered term pairs earlier extracted with *TEA*; If a phrase contains a term only in one language, but not in both, it receives “1” as this case indicates of possible out-of-domain (wrong) translation candidates.
- “2” if a phrase in both languages contained a term pair from the filtered term pairs earlier extracted with *TEA*.

In order to find out whether a phrase contains a given term or not, the phrase and the term itself were stemmed. Finally, the transformed phrase table was integrated back into the adapted SMT system.

Both parts of the ACCURAT improved SMT system (the translation model and the language models) were then tuned with minimum error rate training (MERT) using the same in-domain tuning data that was used for tuning of the baseline system and the intermediate system. The tuning set contained 1,837 unique in-domain sentence pairs. Tuning of the ACCURAT improved SMT system was done in 7 MERT iterations.

5. Evaluation Methodology

The ACCURAT project applies different techniques to evaluate the tools produced in the project. The SMT quality and usability assessment in localization scenario is performed using automatic and human evaluation techniques:

- **Automatic quality score calculation.** This is a standard method in MT evaluation: BLEU, NIST, TER and METEOR scores are calculated to compare the output of the baseline and adapted SMT systems with the human translations from the parallel test corpus.
- **Comparative evaluation.** It compares the output of the adapted system with the output of the baseline SMT system. Translated sentences which are different in the adapted SMT system are compared to sentences translated with the baseline system according to a three-point scale: better–similar–worse quality.
- **Usability for localization:** There are two criteria in this task: text quality (in terms of terminology, style etc.), and translators’ productivity.

6. Automatic evaluation

The evaluation of the baseline and both improved systems was performed through the *LetsMT!* platform with four different automatic evaluation metrics – BLEU, NIST, TER and METEOR. The evaluation set contained 926 unique IT domain sentence pairs. Both case sensitive and case insensitive evaluation variants were produced. The results are given in Table 4.

Table 4. Automatic evaluation results

System	Case sensitive?	BLEU	NIST	TER	METEOR
Baseline	No	11.41	4.0005	85.68	0.1711
	Yes	10.97	3.8617	86.62	0.1203
Intermediate ACCURAT improved system	No	56.28	9.1805	43.23	0.3998
	Yes	54.81	8.9349	45.04	0.3499
ACCURAT improved system	No	56.66	9.1966	43.08	0.4012
	Yes	55.20	8.9674	44.74	0.3514

The automatic evaluation shows a significant performance increase of the improved systems over the baseline system in all evaluation metrics. For the improved systems intermediate results before phrase table transformations show a little increase in performance when making the phrase table term-aware. This is due to better terminology selection in the fully adapted system. As terms comprise only a certain part of texts, the improvement is limited.

7. Comparative evaluation

7.1. Evaluation environment

For comparative evaluation we used Tilde’s web based evaluation environment (Skadiņš et al. 2010) where we can upload source sentences and outputs of two MT systems as simple text (“*.txt”) files. The system’s interface is presented in Figure 4.

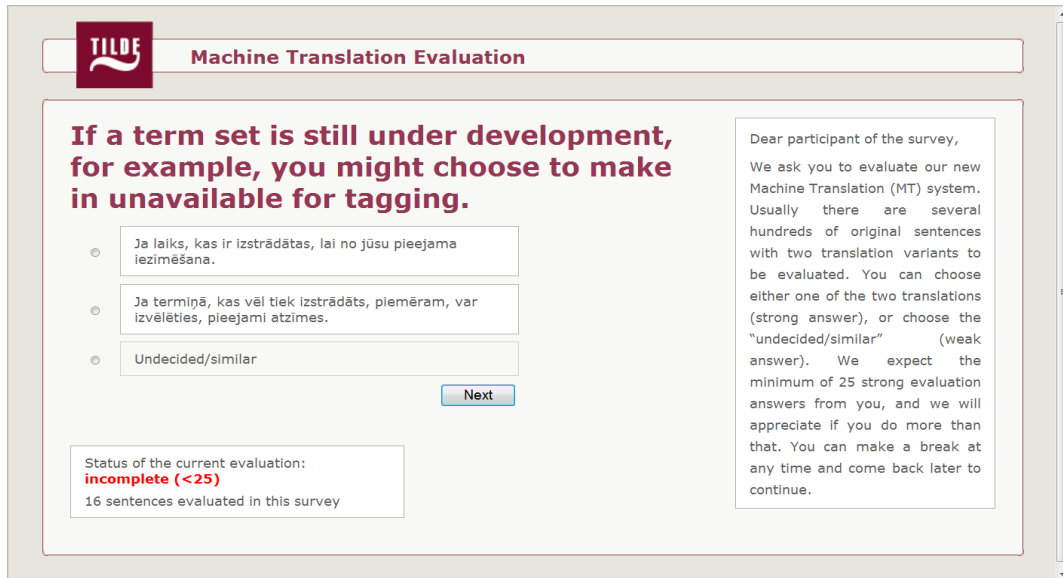


Figure 4 Tilde's web based evaluation environment for the system comparison task

Evaluators are evaluating systems sentence by sentence. A source sentence and output of two SMT systems are shown to evaluators. The order of SMT system outputs is randomized: in some cases the first is an output of the first system, while in other cases the output of the second system comes in the first position. Evaluators are asked to evaluate at least 25 sentences. This can be done in small portions: evaluator can open the evaluation survey and evaluate a few sentences; then go away and come back later to continue with the task.

7.2. Evaluation methodology

We are calculating how often users prefer each system based on all answers and based on comparison of sentences. When we calculate evaluation results based on all answers, we evaluate the percentage from the count of how many times users choose one system to be better than the other using Eq. (1). To be sure about the statistical relevance of results we also calculate confidence interval of the results using Eq. (2).

$$p = \frac{A}{A+B} 100\% \quad (1)$$

$$ci = z \sqrt{\frac{p(1-p)}{A+B}} 100\% \quad (2)$$

where z for a 95% confidence interval is 1.96, A is the number of users preferring the first system, and B is the number of users preferring the second system.

When we have calculated p and ci , then we can say that users prefer the first system over the second in $p \pm ci$ percents of individual evaluations. We say that evaluation results are **weakly sufficient** to say that with a 95% confidence the first system is better than the second if Eq. (3) is true.

$$p - ci > 50\% \quad (3)$$

Such evaluation results are weakly sufficient because they are based on all evaluations but they do not represent system output variations from sentence to sentence.

To get more reliable results we calculated how evaluators have evaluated systems on a sentence level: if we have A evaluators preferring a particular sentence from the first system and B evaluators preferring the sentence from the second system, then we can calculate percentage using Eq. (1) and confidence interval using Eq. (2). We say that a particular

sentence is translated better by the first system than by the other system if Eq. (3) is true. To get more reliable evaluation results we are not asking evaluators to evaluate sentences which have sufficient confidence that they are translated better by one system than by the other. When we have A sentences evaluated to be better translated by the first system and B sentences evaluated to be better translated by the second system or systems are in tie, then we can calculate evaluation results on sentence level using Eqs. (1) and (2) again. And we can say that evaluation results are **strongly sufficient** to say that the first system is better than the second in the sentence level if Eq. (3) is true.

7.3. Evaluation results

For the system comparison we used the same test corpus as for automatic evaluation and compared the baseline system against the ACCURAT improved system. The summary of human evaluation results is presented in Figure 5 (total points) and Figure 6 (count of best sentences), where System 1 is the baseline system and System 2 is the ACCURAT improved system.

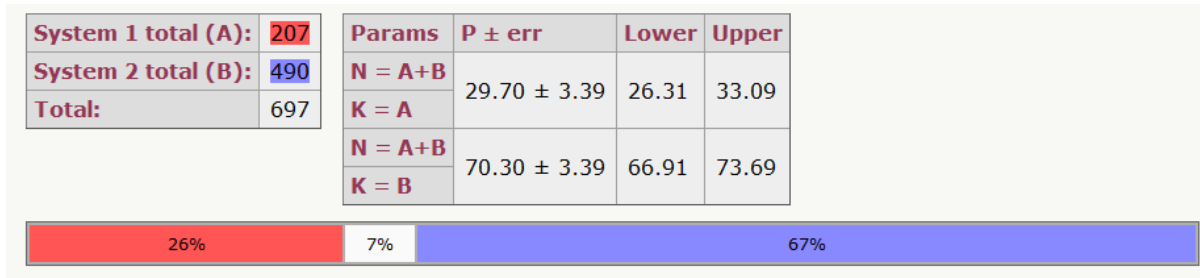


Figure 5. System comparison by total points

The Figure 5 shows that from 697 cases when the sentences were evaluated, in 490 cases (70.30±3.39%) output of the ACCURAT improved SMT system (System 2) was chosen as a better translation, while in 207 cases (29.70±3.39%) users preferred the translation of the baseline system (System 1). This allows us to conclude that adapted SMT system provides better translations as the baseline system for IT domain texts.

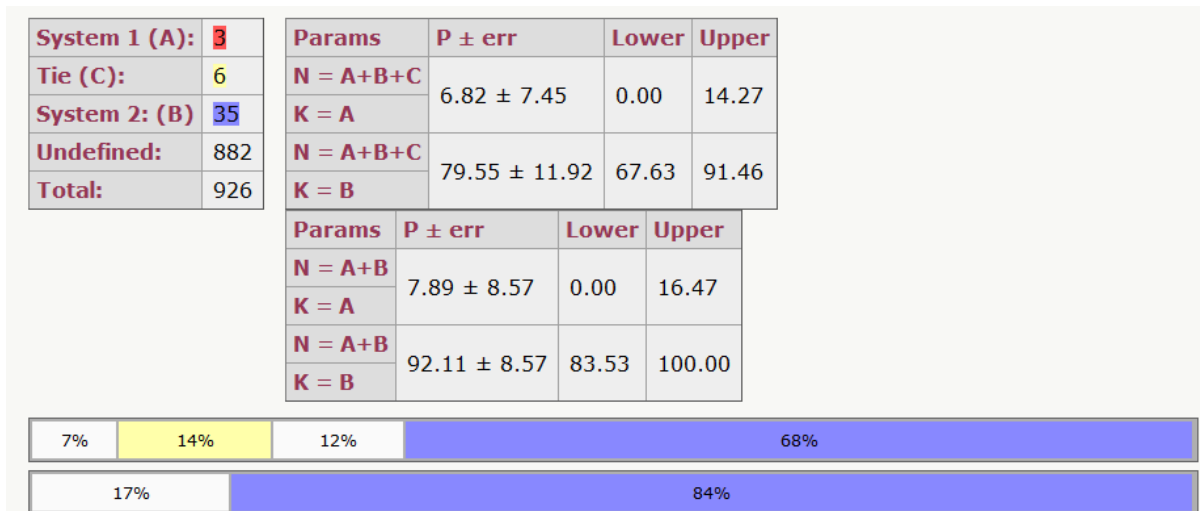


Figure 6. System comparison by count of best sentences

The Figure 6 illustrates evaluation on sentence level: for 35 sentences we can reliably say that the ACCURAT improved SMT system provides a better translation, while only for 3 sentences users preferred the translation of the baseline system. It has to be mentioned here,

that, although in general more sentences were evaluated, here we present results only for sentences, which are reliable.

8. Evaluation in Localization Scenario

Efficiency (translation performance) of translation process without degradation of quality is the most important measure for localization service providers. Thus the main goal of this evaluation task is to evaluate whether integration of ACCURAT results in the localization process allows increasing the efficiency of translation, increasing the output of translators in comparison to the efficiency of manual translation.

8.1. Evaluation methodology

The ACCURAT evaluation in localization scenario is based on the measurement of translation performance calculated as a number of words translated per hour. Translation with the ACCURAT improved SMT system is tested against available manual translation productivity and quality. The productivity increases are evaluated.

8.2. Test data

For tests 30 documents from the software localization domain were used. These documents were split in two parts to perform translation scenarios described below. The length of each part of the document is 250 to 260 adjusted words⁴ in average, resulting in 2 sets of documents with about 7 700 words in each set.

8.3. Evaluation scenarios

Since automatic evaluation and system comparison showed significant improvement in translation quality for the ACCURAT improved SMT system and the quality of the baseline system was rather low (only 11.41 BLEU points), we decided to perform the following two evaluations:

- Translation using translation memories (TM) only.
- Translation offering machine translation suggestions of translation memories and the SMT system that is enriched with ACCURAT data.

The localisation workflows of the two scenarios are given in Figure 7.

⁴ An adjusted word is a metric used for quantifying work to be done by translators.

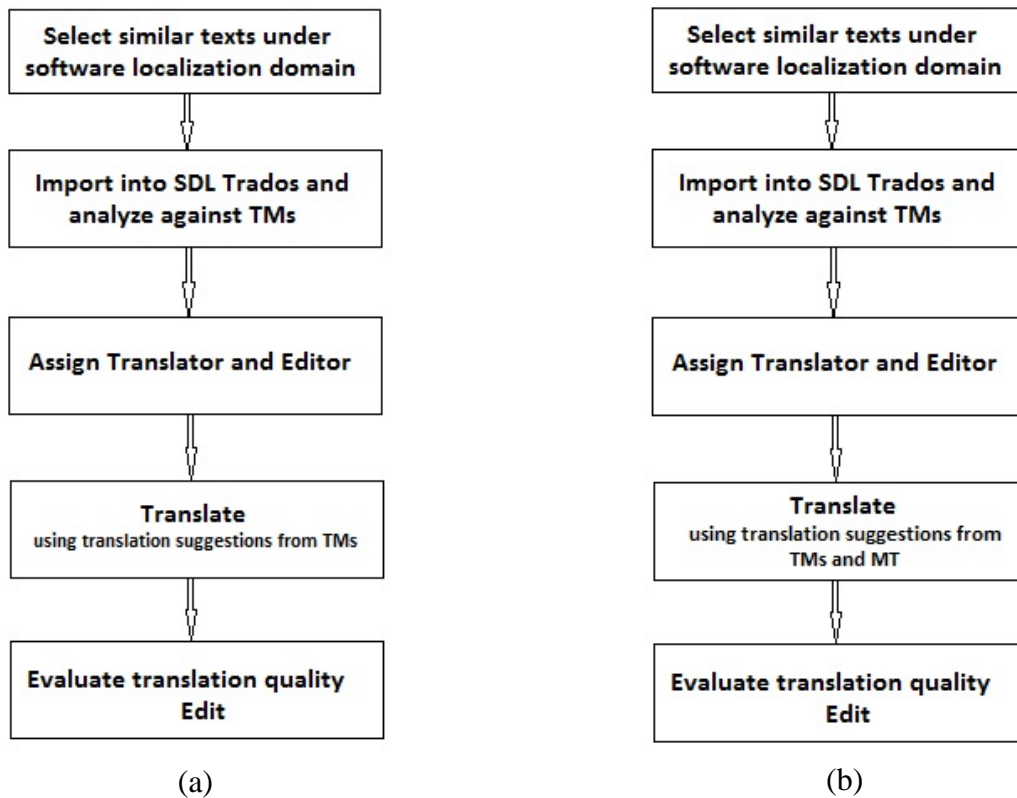


Figure 7. Localisation business workflows without (a) and with (b) integrated SMT support

8.4. Tools: SDL Trados + SMT

For the second evaluation scenario the ACCURAT improved SMT system was integrated into the SDL Trados 2009 CAT environment. The LetsMT! platform provides a plug-in for the SDL Trados 2009 CAT environment to use generated MT systems. The ACCURAT SMT system is running on the LetsMT! platform and is accessible using a web service interface based on the SOAP protocol. The ACCURAT improved SMT system is used to provide translation recommendations for those translation segments that do not have exact matches or close matches in the translation memories. Suggestions that come from the MT system are clearly marked. Localization specialists will be able to choose these translations and post-edit them to create a professional translation.

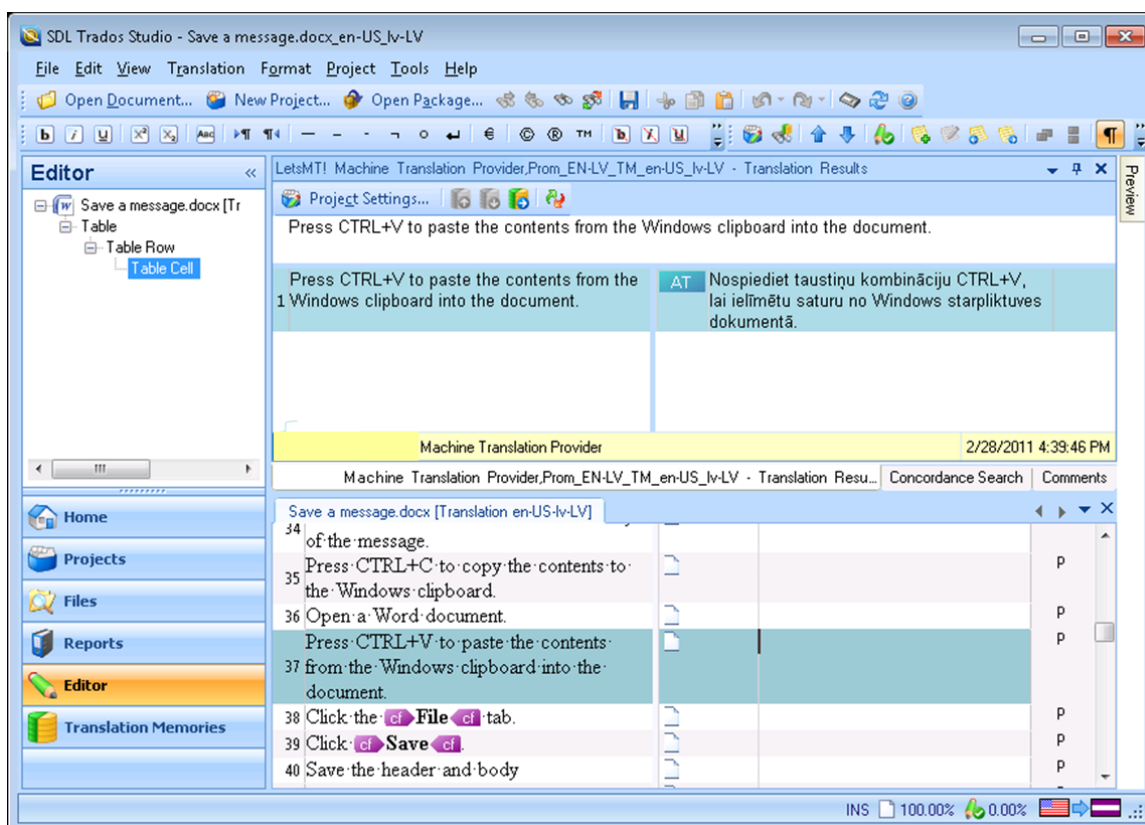


Figure 8. Interface of the SDL Trados Studio with integrated MT output

8.5. Evaluation and results

8.5.1. Test execution

Three translators with different levels of experience and average performance were involved in the evaluation cycle. Altogether 60 documents (30 without SMT and 30 with) were translated using the SDL Trados Studio platform. Every document was entered in the translation project tracking system as a separate translation task. Each of the two evaluators had to translate 10 documents without SMT support and 10 documents with integrated SMT support.

8.5.2. Evaluation procedure

After the whole text of a document is translated by a translator, it is evaluated for translation performance and translation quality by Editors. Editors have no information about techniques used to assist translators.

Quality of work is measured by filling a QA checklist in accordance to Tilde's QA process (Appendix 2). The following text quality areas are measured: accuracy, language quality, style and terminology.

Performance and quality of work in both evaluation scenarios is measured and compared for every individual translator. Individual productivity of each translator is measured and compared against his or her own productivity. An error score will be calculated for every translation task. The error score is a metric calculated by counting errors identified by the editor and applying a weighted multiplier based on the severity of the error type.

The error score is calculated per 1000 words and it is calculated as:

$$ErrorScore = \frac{1000}{n} \sum_i w_i e_i$$

where

n is a number of words in a translated text,

e_i is a number of errors of type i ,

w_i is a coefficient (weight) indicating severity of type i errors.

There are 15 different error types grouped in 4 error classes – accuracy, language quality, style and terminology. Different error types influence the error score differently because errors have a different weight depending on the severity of the error type. For example, errors of the type comprehensibility (an error that obstructs the user from understanding the information; very clumsy expressions) have weight 3, while errors of type omissions/unnecessary additions have weight 2.

Depending on the error score the translation is assigned a translation quality grade: *Superior*, *Good*, *Mediocre*, *Poor* and *Very poor* (Table 5).

Table 5. Quality evaluation based on the score of weighted errors

Error Score	Quality Grade
0...9	Superior
10...29	Good
30...49	Mediocre
50...69	Poor
>70	Very poor

8.5.3. Evaluation results

The results were analysed for 60 translation tasks (30 tasks in each scenario) by analysing average values for translation performance (translated words per hour) and an error score for translated texts.

Usage of MT suggestions in addition to translation memories increased productivity of the translators in average from 503 to 572 words per hour (13.6% improvement). There were significant differences in the results of different translators from performance increase by 35.4% to decreased performance by 5.9% for one of the translators (see Table 6). Analysis of these differences requires further studies but most likely they are caused by working patterns and the skills of individual translators. The average productivity for all translators on the localisation task has been calculated using the following formula:

$$Productivity (scenario) = \frac{\sum_{Text=1}^N Adjusted\ words(Text, scenario)}{\sum_{Text=1}^N Actual\ time(Text, scenario)}$$

Table 6 Localisation scenario productivity evaluation results

Translator	Scenario	Actual productivity	Productivity increase or decrease
Translator 1	S1	493.2	35.39%
	S2	667.7	
Translator 2	S1	380.7	13.02%
	S2	430.3	
Translator 3	S1	756.9	-5.89%
	S2	712.3	
All	S1	503.2	13.63%
	S2	571.9	

According to the standard deviation of productivity in both scenarios (without MT support 186.8 and with MT support 184.0) there were no significant performance differences in the overall evaluation (see Table 7). However, each translator separately showed higher differences in translation performance when using the MT translation scenario.

Table 7. Standard deviation of productivity

Translator	Scenario	Standard deviation of productivity
Translator 1	S1	110.7
	S2	121.8
Translator 2	S1	34.2
	S2	38.9
Translator 3	S1	113.8
	S2	172.0
All	S1	186.8
	S2	184.0

The overall error score (shown in Table 8) increased for one out of three translators. Although the total increase in the error score for all translators combined was from 24.9 to 26.0 points, it still remained at the quality evaluation grade “Good”.

Table 8. Localisation task error score results

Translator	Scenario	Accuracy	Language quality	Style	Terminology	Total error score
Translator 1	S1	6.8	8.0	6.8	1.6	23.3
	S2	9.9	14.4	7.8	4.1	36.3
Translator 2	S1	8.2	10.1	11.7	0.0	30.0
	S2	3.8	11.7	7.6	1.5	24.6
Translator 3	S1	4.6	9.5	7.3	0.0	21.4
	S2	3.0	8.3	6.0	0.8	18.1
All	S1	6.5	9.3	8.6	0.5	24.9
	S2	5.4	11.4	7.1	2.1	26.0

Detailed results of the English-Latvian localisation scenario are given in Appendix 3.

Conclusion

To our knowledge this is the first evaluation of usability of SMT system enriched with comparable data for particular domain in the localization environment for translation into a less-resourced highly inflected language. This is also one of the first evaluations of SMT for a less-resourced highly inflected language in the localization environment.

The results of our experiment clearly demonstrate that it is feasible to adapt SMT systems for a particular domain with the help of comparable data and integrate such SMT systems for highly inflected languages into the localization process.

The use of the English->Latvian domain adapted SMT suggestions (trained on comparable data) in addition to the translation memories in the SDL Trados CAT tool leads to the increase of translation performance by 13.6% while maintaining an acceptable (“*Good*”) quality of the translation. However, our experiments also showed a relatively high difference in translator performance changes (from -5.89% to +35.39%), which suggests that for more justified results the experiment should be carried out with more than three participants. Such an experiment would also be useful for analysis of how translator average performance levels without MT support impact the possible increase of performance after adding MT support.

Error rate analysis shows that overall usage of MT suggestions decrease the quality of the translation in two error categories (language quality and terminology). At the same time this degradation is not critical and the result is acceptable for production purposes.

References

- Flournoy, R. and Duran C. 2009. Machine translation and document localization at Adobe: From pilot to production. In: MT Summit XII: proceedings of the twelfth Machine Translation Summit, Ottawa, Canada.
- Kelly, N., DePalma, D.A., and Stewart, R.G., 2012. The Language Service Market 2012, Common Sense Advisory, 2012, Lowell, USA
- Koehn, P., Federico M., Cowan B., Zens R., Duer C., Bojar O., Constantin A., Herbst E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, 177-180.
- O'Brien, S. 2005. Methodologies for measuring the correlations between post-editing effort and machine translatability. In: Machine Translation, 19(1):37–58, March 2005.
- Papineni, K., Roukos, S., Ward, T., Zhu, W. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of ACL, Philadelphia.
- Pinnis, M., Ion, R., Ștefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A. and Babych, B. 2012a. ACCURAT Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora. In: Proceedings of System Demonstrations Track of ACL 2012, Jeju Island, Republic of Korea, 8-14 July, 2012.
- Pinnis, M., Ljubešić, N., Ștefănescu, D., Skadiņa, I., Tadić, M. and Gornostay, T. 2012b. Term Extraction, Tagging and Mapping Tools for Under-Resourced Languages. In: Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012), Madrid, Spain.
- Plitt, M., Masselot, Fr. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. In: The Prague Bulletin of Mathematical Linguistics 93, p.7-16.
- Schmidtke, D. 2008. Microsoft office localization: use of language and translation technology. *URL* <http://www.tm-europe.org/files/resources/TM-Europe2008-Dag-Schmidtke-Microsoft.pdf>.
- Skadiņš R., Goba K. and Šics V. 2010. Improving SMT for Baltic Languages with Factored Models. In: Proceedings of the Fourth International Conference Baltic HLT 2010, Riga.
- Skadiņš R., Puriņš M., Skadiņa I., Vasiļjevs A. 2011. Evaluation of SMT in localization to under-resourced inflected language. In: Proceedings of the 15th International Conference of the European Association for Machine Translation EAMT 2011, 35-40, May 30-31, 2011, Leuven, Belgium.
- Vasiļjevs, A., Gornostay, T. and Skadins, R. 2010. LetsMT! – Online Platform for Sharing Training Data and Building User Tailored Machine Translation. In: Proceedings of the Fourth International Conference Baltic HLT 2010, Riga.
- Wintergreen Research, 2011. Language Translation Software Market Shares and Forecasts, Worldwide, 2011-2017, 2011, Lexington.

Appendix 1. Creation of English-Latvian comparable corpora

This appendix provides a detailed description of how the English-Latvian software localisation domain comparable corpora were created.

We have used existing comparable corpora and created new artificial comparable corpora: (1) comparable corpora collected by ILSP using the focused monolingual crawler (*FMC*), (2) comparable corpora from different version software manuals, and (3) comparable corpus from artificially polluted parallel data.

1. IT domain comparable corpus collected with focused monolingual crawler (FMC)

The first corpus we used was the weakly comparable IT domain corpus collected from the Web by ILSP using the *FMC* crawler described in the Deliverable 3.5 *Tools for building comparable corpus from the Web*. The corpus was collected using seed lists of terms (112 terms in English and Latvian) and seed URL list (80 URLs for Latvian and 39 URLs for English). The statistics of the *FMC* corpus are given below:

Parameter	English	Latvian
Documents	7,722	1,085
Unique sentences	232,665	96,573
Tokens in unique sentences	4,369,457	1,580,352

As it was mentioned before, this corpus can be characterized as weakly comparable, from which a rather small amount (about 1000 sentence pairs) of pseudo-parallel data can be extracted. Therefore this corpus is used for two purposes: (1) monolingual data resource for language modelling and (2) in-domain non-comparable data resource for parallel data pollution in order to create the third strongly comparable software localisation domain corpus.

2. Comparable corpora from different version software manuals

Software manuals from different versions of the same application are often strongly comparable. Such corpora may also be an important source for parallel sentences. Using proprietary translation memories containing different versions of software manuals we created three artificial comparable corpora:

- The first two corpora we obtained by splitting the translation memories of the different software versions in smaller files of up to 100 paragraphs and aligning them with the *DictMetric* comparability metric (developed by CTS), so that *DictMetric* compares documents only from different versions of the same software (for instance, software manuals of “*version 1*” in one language and software manuals of “*version 2*” in another language).
- For the third corpora we split the translation memories with a different approach: for each 200 paragraphs in a translation memory, the first 100 paragraphs of the TMs were taken from one language and the remaining 100 paragraphs were taken from the other language. The monolingual corpora were then aligned at the document level with *DictMetric*. This procedure will further show whether there has been parallel data overlap in the translation memories of the same software manuals.

For all three corpora a *DictMetric* comparability score threshold of 0.3 was applied. The monolingual corpora statistics are shown below:

Corpus	Parameter	English	Latvian
SW Manual 1	Documents	2,288	1,708
	Unique sentences	197,537	179,547
	Tokens in unique sentences	3,300,105	1,390,451
SW Manual 2	Documents	1,681	2,297
	Unique sentences	179,234	201,031
	Tokens in unique sentences	1,569,129	2,857,590
SW Manual 3	Documents	1,231	1,231
	Unique sentences	111,981	113,352
	Tokens in unique sentences	1,825,171	1,582,286

Because of the corpora being in a narrow domain, *DictMetric* produced many document pairs over 0.3. In order to acquire a reasonable number of aligned documents, we filtered *DictMetric* results so that for each source and target language document there would be no more than the top three alignments. This procedure created two lists of document pairs (top 3 Latvian documents for each English document and top 3 English documents for each Latvian document), which were combined to create the final comparable corpora (aligned in document level). The comparable corpora statistics are shown below:

Corpus	English documents	Latvian documents	Number of aligned document pairs	Number of aligned document pairs after filtering
SW Manual 1	2,288	1,708	124,795	8,930
SW Manual 2	1,681	2,297	156,449	9,155
SW Manual 3	1,231	1,231	81,832	5,314

3. Comparable corpus from artificially polluted parallel data

Since the collected IT domain Web corpora described in point 1 was weakly comparable and the amount of extracted data was insufficient for creation of an in-domain SMT system useful for localization purposes, we used a proprietary parallel corpus in the software localisation domain containing 1,257,142 parallel sentence pairs and created an artificial comparable corpus out of it by polluting it with in-domain non-comparable (because of random selection) data from the corpus collected with *FMC* tool. We name this corpus “*SW Mixed*”. By this we want to show what can be achieved with ACCURAT methods if such large comparable corpus exists, that is, we want to show that for under-resourced MT domains with enough comparable data the ACCURAT methods can be beneficial localization tasks.

The pseudo-code of the corpora creation process is as follows:

Read a parallel corpus and weakly comparable in-domain corpora into memory (for instance, three list type data structures).

While there are parallel sentences left in the parallel data do the following:

pick a random number between 40 and 70;

open two output file streams (one for the source language and one for the target language text) using a file counter for file name indexing;

while the parallel sentence count in the source and target files is smaller than the generated random number do the following:

write a random number (from 0 to 3) of randomly selected source language non-comparable in-domain sentences within the source language file;

write a random number (from 0 to 3) of randomly selected target language non-comparable in-domain sentences within the target language file;

write a random number (from 1 to 5) of randomly selected parallel sentence pairs within the source and target language files;

remove the written parallel sentence pairs from the remaining parallel data list;

write a random number (from 0 to 3) of randomly selected source language non-comparable in-domain sentences within the source language file;

write a random number (from 0 to 3) of randomly selected target language non-comparable in-domain sentences within the target language file;
close the output file streams.

Following this pseudo-code the resulting corpus contains aligned document pairs that are from weakly to strongly comparable (in very, very rare cases also parallel). With this technique we can also test how well ACCURAT methods can find parallel data within comparable corpora. The monolingual corpora statistics of the in-domain comparable corpora are given below:

Parameter	English	Latvian
Documents	22,498	22,498
Unique sentences	1,316,764	1,215,019
Tokens in unique sentences	16,927,452	13,036,066

Appendix 2. Tilde Translation Quality Assessment Form

This form is filled out by an Editor or a Language Specialist.

Please see procedural notes and description of error categories in **Error categories** sheet.

Fill in the **Basic information** section, **Amount of errors** column and **General comment** field.

Basic information	
Project name:	
File name:	
Source language:	
Target language:	
Translator:	
Validated by:	
Validation date:	
Stylistic type (please, select):	
Number of words checked:	1000

Error Category	Weight	Amount of errors	Negative points
1. Accuracy			
1.1. Understanding of the source text	3		0
1.2. Understanding the functionality of the product	3		0
1.3. Comprehensibility	3		0
1.4. Omissions/Unnecessary additions	2		0
1.5. Translated/Untranslated	1		0
1.6. Left-overs	1		0
Total			0
2. Language quality			
2.1. Grammar	2		0
2.2. Punctuation	1		0
2.3. Spelling	1		0
Total			0
3. Style			
3.1. Word order, word-for-word translation	1		0
3.2. Vocabulary and style choice	1		0
3.3. Style Guide adherence	2		0
3.4. Country standards	1		0
Total			0
4. Terminology			
4.1. Glossary adherence	2		0
4.2. Consistency	2		0
Total			0
Grand Total			0
Error Score (negative points) per 1000 words			0
Quality:			Superior

General comment:

--

Final assessment is done as follows:	Score scale	
Negative points for errors of each category are calculated according to the formula: "Number of errors of given type" x "Error weight"	Error score	Quality grade
Weighted score is calculated according to the following formula: (Total negative points / Wordcount) x 1000	0...9	Superior
Final quality assessment is done according to the Score Scale .	10...29	Good
	30...49	Mediocre
	50...69	Poor
	70...	Very poor

Notes:

In case of recurring errors (double space, the same spelling or terminology error) they should only be counted once.

Each error is counted once, by the most appropriate category. If in doubt, use the first appropriate category (top-down).

Preferential changes should not be counted as negative points, but they may be listed in a separate Comments section.

Category	Description
Accuracy	
Understanding of the source text	A lack of comprehension of the source text resulting in incorrect meaning of the translation.
Understanding the functionality of the product	Translation does not comply with the actual function of the product. The translation of the word is OK as such but incorrect in the context.
Comprehensibility	Any error that obstructs the user from understanding the information. Very clumsy expressions.
Omissions/unnecessary additions	Words, part of sentences, sentences, paragraphs are missing. No relevant information in the source language should be omitted in the translation, unless specifically requested. The translation should not contain any unnecessary text.
Translated/Untranslated	Parts that were supposed to be translated were not translated or parts that should not be translated were translated.
Left-overs	Redundant words resulting from sentence change, wrong declinations resulting from correcting one word only but not the rest. Unnecessary question marks or asterisks left in translated text.
Language quality	
Grammar	Grammar, syntax or morphology rules are broken.
Punctuation	Incorrect usage of punctuation marks - full stops missing, opening or closing punctuation marks (quote, parenthesis), double spaces, etc.
Spelling	The translation should contain no spelling errors.
Style	
Word order, word-for-word translation	Functional sentence perspective (theme, rheme), word order. Word for word translation, resulting in stylistically inappropriate expression.
Vocabulary and style choice	Archaisms, jargon, colloquial words, verbosity, inappropriate style.
Style Guide adherence	Product Style Guide rules are ignored. In case of absence of Product Style Guide definite company style rules must be observed. Standard phrases must be used - in case of technical documentation.
Country standards	Adaptation of country standards (date and time formats, units of measurement, currency, number formats, sorting order, capitalization etc.). Examples (of names, streets, etc.) are not localized.
Terminology	
Glossary adherence	Translation does not adhere to the terms in the glossary of project/product, or does not use generally available industry terminology. Technical documentation does not use the correct translation of interface elements.
Consistency	Inconsistent usage of translation for one term or title (for cross-references).

Quality Assessment form, Values for form fields

Yes/No	Yes No
Languages	English Estonian Latvian Lithuanian
Text Type	User interface User assistance, tech. documentation Medicine Legal Marketing or Web material
Quality	Superior Good Mediocre Poor Very poor
Error category	Accuracy Language quality Style Terminology Preferential

Appendix 3. Detailed evaluation results in localisation scenario

This appendix provides detailed results of individual translators participating in evaluation of ACCURAT improved English-Latvian MT system in localisation scenario.

Task ID (file name in LPS)	Scenario (S1, S2)	Text size (adjusted words)	Translator name	Translator qualification	Estimated time (h)	Planned performance (adjusted words/h)	Actual time (h)	Actual performance (adjusted words/h)	Quality assessment, negative points						Quality total valuation (Superior, Good, Mediocre, Poor, Very Poor)
									Accuracy	Language quality	Style	Terminology	Total	Total (per 1000 words)	
Text 10-1	S1	252	Translator 1	Senior Translator	0.72	350	0.5	504	0	0	1	0	1	4	Superior
Text 1-1	S1	316.6	Translator 2	Translator	0.9	350	0.95	333	7	5	3	0	15	47	Mediocre
Text 11-1	S1	208.4	Translator 2	Translator	0.6	350	0.6	347	0	3	2	0	5	24	Good
Text 12-1	S1	270	Translator 3	Senior Translator	0.77	350	0.4	675	0	1	2	0	3	11	Good
Text 13-1	S1	280	Translator 2	Translator	0.8	350	0.7	400	2	6	1	0	9	32	Mediocre
Text 14-1	S1	280	Translator 1	Senior Translator	0.8	350	0.38	737	3	0	1	2	6	21	Good
Text 15-1	S1	254	Translator 2	Translator	0.73	350	0.65	391	0	2	4	0	6	24	Good
Text 16-1	S1	231	Translator 1	Senior Translator	0.66	350	0.5	462	6	0	3	0	9	39	Mediocre
Text 17-1	S1	294	Translator 1	Senior Translator	0.84	350	0.66	445	0	6	2	0	8	27	Good
Text 18-1	S1	188	Translator 2	Translator	0.54	350	0.45	418	3	0	2	0	5	27	Good
Text 19-1	S1	281	Translator 3	Senior Translator	0.8	350	0.42	669	0	6	3	0	9	32	Mediocre
Text 20-1	S1	218	Translator 3	Senior Translator	0.62	350	0.28	779	0	4	4	0	8	37	Mediocre
Text 2-1	S1	228	Translator 3	Senior Translator	0.65	350	0.38	600	0	4	1	0	5	22	Good
Text 21-1	S1	210.7	Translator 1	Senior Translator	0.6	350	0.5	421	3	0	2	0	5	24	Good
Text 22-1	S1	283	Translator 3	Senior Translator	0.77	350	0.34	832	3	5	1	0	9	32	Mediocre
Text 23-1	S1	236	Translator 3	Senior Translator	0.67	350	0.24	983	3	2	1	0	6	25	Good

Task ID (file name in LPS)	Scenario (S1, S2)	Text size (adjusted words)	Translator name	Translator qualification	Estimated time (h)	Planned performance (adjusted words/h)	Actual time (h)	Actual performance (adjusted words/h)	Quality assessment, negative points						Quality total valuation (Superior, Good, Mediocre, Poor, Very Poor)
									Accuracy	Language quality	Style	Terminology	Total	Total (per 1000 words)	
Text 24-1	S1	281	Translator 3	Senior Translator	0.8	350	0.3	937	0	1	2	0	3	11	Good
Text 25-1	S1	270.7	Translator 1	Senior Translator	0.77	350	0.66	410	0	1	1	2	4	15	Good
Text 26-1	S1	207.9	Translator 2	Translator	0.59	350	0.5	416	2	0	2	0	4	19	Good
Text 27-1	S1	168.5	Translator 1	Senior Translator	0.48	350	0.5	337	0	2	2	0	4	24	Good
Text 28-1	S1	231.1	Translator 1	Senior Translator	0.64	350	0.42	550	3	3	2	0	8	35	Mediocre
Text 29-1	S1	291.3	Translator 1	Senior Translator	0.83	350	0.5	583	0	4	3	0	7	24	Good
Text 30-1	S1	310.9	Translator 2	Translator	0.89	350	0.7	444	5	3	2	0	10	32	Mediocre
Text 3-1	S1	298	Translator 3	Senior Translator	0.85	350	0.38	784	0	1	3	0	4	13	Good
Text 4-1	S1	275	Translator 2	Translator	0.79	350	0.8	344	2	0	1	0	3	11	Good
Text 5-1	S1	256.4	Translator 1	Senior Translator	0.75	350	0.42	610	2	4	0	0	6	23	Good
Text 6-1	S1	265	Translator 2	Translator	0.76	350	0.7	379	0	3	8	0	11	42	Mediocre
Text 7-1	S1	270	Translator 3	Senior Translator	0.77	350	0.38	711	3	0	2	0	5	19	Good
Text 8-1	S1	264	Translator 2	Translator	0.75	350	0.7	377	0	4	5	0	9	34	Mediocre
Text 9-1	S1	254	Translator 3	Senior Translator	0.73	350	0.34	747	3	1	0	0	4	16	Good
Text 10-2	S2	245.1	Translator 1	Senior Translator	0.7	350	0.5	490	6	0	0	2	8	33	Mediocre
Text 11-2	S2	273.6	Translator 2	Translator	0.78	350	0.65	421	0	4	2	0	6	22	Good
Text 1-2	S2	316.4	Translator 2	Translator	0.9	350	0.75	422	5	4	0	2	11	35	Mediocre
Text 12-2	S2	279.4	Translator 3	Senior Translator	0.8	350	0.5	559	3	3	0	0	6	21	Good
Text 13-2	S2	274.1	Translator 2	Translator	0.78	350	0.65	422	0	0	0	0	0	0	Superior
Text 14-2	S2	250.8	Translator 1	Senior Translator	0.72	350	0.4	627	4	0	2	4	10	40	Mediocre

Task ID (file name in LPS)	Scenario (S1, S2)	Text size (adjusted words)	Translator name	Translator qualification	Estimated time (h)	Planned performance (adjusted words/h)	Actual time (h)	Actual performance (adjusted words/h)	Quality assessment, negative points						Quality total valuation (Superior, Good, Mediocre, Poor, Very Poor)
									Accuracy	Language quality	Style	Terminology	Total	Total (per 1000 words)	
Text 15-2	S2	244	Translator 2	Translator	0.7	350	0.55	444	2	6	2	0	10	41	Mediocre
Text 16-2	S2	292	Translator 1	Senior Translator	0.83	350	0.53	551	0	4	2	0	6	21	Good
Text 17-2	S2	185	Translator 1	Senior Translator	0.53	350	0.25	740	0	4	3	2	9	49	Mediocre
Text 18-2	S2	245.1	Translator 2	Translator	0.7	350	0.5	490	0	2	3	2	7	29	Good
Text 19-2	S2	216	Translator 3	Senior Translator	0.62	350	0.4	540	0	1	2	0	5	23	Good
Text 20-2	S2	302	Translator 3	Senior Translator	0.86	350	0.4	755	0	1	0	0	1	3	Superior
Text 21-2	S2	213.2	Translator 1	Senior Translator	0.61	350	0.3	711	0	1	2	0	3	14	Good
Text 2-2	S2	331.4	Translator 3	Senior Translator	0.81	350	0.3	1105	1	2	1	2	6	18	Good
Text 22-2	S2	270.9	Translator 3	Senior Translator	0.95	350	0.54	502	2	3	1	0	6	22	Good
Text 23-2	S2	283	Translator 3	Senior Translator	0.81	350	0.34	832	0	5	3	0	8	28	Good
Text 24-2	S2	229.2	Translator 3	Senior Translator	0.65	350	0.3	764	0	2	3	0	5	22	Good
Text 25-2	S2	234	Translator 1	Senior Translator	0.67	350	0.3	780	0	1	2	0	3	13	Good
Text 26-2	S2	298.4	Translator 2	Translator	0.85	350	0.6	497	0	2	1	0	5	5	Superior
Text 27-2	S2	230.4	Translator 1	Senior Translator	0.66	350	0.3	768	5	2	1	2	10	43	Mediocre
Text 28-2	S2	274.1	Translator 1	Senior Translator	0.78	350	0.42	653	0	7	2	0	11	40	Mediocre
Text 29-2	S2	237.2	Translator 1	Senior Translator	0.68	350	0.25	949	9	6	5	0	20	84	Very poor
Text 30-2	S2	257.7	Translator 2	Translator	0.74	350	0.55	469	3	7	0	0	10	39	Mediocre
Text 3-2	S2	271	Translator 3	Senior Translator	0.77	350	0.38	713	0	2	4	0	6	22	Good
Text 4-2	S2	215	Translator 2	Translator	0.61	350	0.55	391	0	0	3	0	3	14	Good
Text 5-2	S2	262.1	Translator 1	Senior Translator	0.75	350	0.38	690	0	10	0	0	12	46	Mediocre

Task ID (file name in LPS)	Scenario (S1, S2)	Text size (adjusted words)	Translator name	Translator qualification	Estimated time (h)	Planned performance (adjusted words/h)	Actual time (h)	Actual performance (adjusted words/h)	Quality assessment, negative points						Quality total valuation (Superior, Good, Mediocre, Poor, Very Poor)
									Accuracy	Language quality	Style	Terminology	Total	Total (per 1000 words)	
Text 6-2	S2	301	Translator 2	Translator	0.86	350	0.8	376	0	2	4	0	6	20	Good
Text 7-2	S2	200	Translator 3	Senior Translator	0.57	350	0.26	769	0	1	1	0	2	10	Good
Text 8-2	S2	221	Translator 2	Translator	0.63	350	0.55	402	0	4	5	0	9	41	Mediocre
Text 9-2	S2	267	Translator 3	Senior Translator	0.76	350	0.3	890	2	2	1	0	5	19	Good